# Améliorer l'intelligibilité des arbres de décision avec des explications probabilistes concises et fiables

Louenas Bounia*

Univ. Artois, CNRS, CRIL, F-62300 Lens*
nom@cril.fr,
http://www.cril.univ-artois.fr/

**Résumé.** Ce travail traite de l'intelligence artificielle explicable (IA explicable), en particulier l'amélioration de l'intelligibilité des arbres de décision à l'aide des explications probabilistes fiables et concises. Les arbres de décision sont populaires car ils sont considérés comme hautement interprétables. En raison des limitations cognitives, les explications abductives peuvent être de trop grandes taille pour être interprétables par les utilisateurs humains. Lorsque cela se produit, les arbres de décision sont loin d'être facilement interprétables. Dans ce contexte, notre objectif est d'améliorer l'intelligibilité des arbres de décision en utilisant des explications probabilistes. En nous inspirant d'un précédent travail sur l'approximation des explications probabilistes, nous proposons un algorithme glouton qui permet de dériver des explications probabilistes concises et fiables pour les arbres de décision. Nous décrivons en détail cet algorithme et le comparant à l'encodage SAT de l'état de l'art, en mettant en avant le gain en intelligibilité et en soulignant son efficacité empirique.

## 1 Introduction

**Contexte.** Expliquer une décision à une personne consiste à fournir des détails ou des raisons qui l'aident à comprendre pourquoi la décision a été prise. C'est particulièrement important lorsque les décisions sont prises par des modèles d'apprentissage automatique (ML), tels que les arbres de décision et les forêts aléatoires (Breiman, 2001), les réseaux de Markov, les machines à vecteurs de support et les réseaux de neurones profonds. À mesure que le nombre d'applications qui reposent sur les techniques de ML augmente, la recherche sur l'IA explicable est devenue de plus en plus importante ces dernières années. Les approches XAI peuvent généralement être catégorisées comme des méthodes indépendantes du modèle dites "agnostique". Par exemple, LIME (Ribeiro et al., 2016), SHAP (Lundberg et Lee, 2017) et Anchors (Ribeiro et al., 2018) ou des approches formelles qui fournissent des explications abductives et des raisons suffisantes (Shih et al., 2018). En effet, il a été rapporté par (Ignatiev et al., 2019) qu'une explication $t$ peut être cohérente avec différentes classes prédites.

Une limitation importante des approches XAI formelles est que les explications peuvent être de grande taille. Il est important de garder à l'esprit que l'explication est un processus social (Miller, 2019; Molnar, 2020), où les utilisateurs sont des êtres humains qui ont des

limitations cognitives inhérentes. Dans son article fondateur (Miller, 1956), le psychologue G. Miller a introduit l'idée de "chunking" des objets par les gens (c'est-à-dire de les regrouper en une unité) et a affirmé que, en raison des limitations de la mémoire humaine, la taille des chunks est limitée à $7\pm2$. Un travail récent a étudié les explications probabilistes comme mécanisme pour réduire la taille des explications et les rendre plus concises et adaptées à l'explication du monde réel (Wäldchen et al., 2021). Le problème de décision consiste à vérifier si une instance $x$ admet une raison $\delta$-probable de taille $k$ sous une fonction booléenne $f$, où $f$ est spécifiée comme une formule CNF, est $\mathsf{NP}^{\mathsf{PP}}$-complet (Wäldchen et al., 2021). Ce résultat montre que le problème est très difficile.

**Objectif.** Notre objectif est d'améliorer l'intelligibilité des arbres de décision en utilisant des explications probabilistes. Nous nous sommes inspirés d'un travail précédent sur l'approximation des explications probabilistes (Bounia et Koriche, 2023). Ce travail constitue une application directe des résultats obtenus dans (Bounia et Koriche, 2023). En nous basant sur ces résultats, nous proposons un algorithme glouton permettant de générer des explications probabilistes concises et fiables pour les arbres de décision. Nous décrivons en détail cet algorithme et le comparons empiriquement à l'encodage SAT de l'état de l'art (Arenas et al., 2022), mettant en évidence l'amélioration de l'intelligibilité et soulignant son efficacité.

# 2 Arbre de décision et DNF orthogonale

## 2.1 Préliminaires

Pour un entier $n$, soit $[n]$ l'ensemble $\{1, \ldots, n\}$. On note $\mathcal{F}_n$ la classe de toutes les fonctions booléennes de $\{0,1\}^n$ à $\{0,1\}$, et on utilise $X_n = \{x_1, \ldots, x_n\}$ pour désigner l'ensemble des variables booléennes. Toute affectation $\boldsymbol{x} \in \{0,1\}^n$ est appelée une *instance*. Un *littéral* $\ell$ est une variable $x_i$ ou sa négation $\neg x_i$, également notée $\overline{x}_i$. $x_i$ et $\overline{x}_i$ sont des littéraux complémentaires. Un *terme* $t$ est une conjonction de littéraux, et une *clause* $c$ est une disjonction de littéraux. $Lit(f)$ désigne l'ensemble de tous les littéraux de $f$. Une formule DNF est une disjonction de termes et une formule CNF est une conjonction de clauses. L'ensemble des variables apparaissant dans une formule $f$ est noté $Var(f)$. Une formule $f$ est *consistante* si et seulement si elle a un modèle. Une formule CNF est *monotone* lorsque chaque littéral d'une variable donnée dans la formule a la même polarité. Une formule $f_1$ *implique* une formule $f_2$, notée $f_1 \models f_2$, si et seulement si chaque modèle de $f_1$ est un modèle de $f_2$. Deux formules $f_1$ et $f_2$ sont *équivalentes*, noté $f_1 \equiv f_2$, si et seulement si elles ont les mêmes modèles. Étant donné une assignation $z \in \{0,1\}^n$, le terme correspondant est défini comme

$$t_z = \bigwedge_{i=1}^{n} x_i^{z_i} \text{ où } x_i^0 = \overline{x}_i \text{ et } x_i^1 = x_i$$

Un terme $t$ *couvre* une affectation $z$ si $t \subseteq t_z$. Un *impliquant* d'une fonction booléenne $f$ est un terme qui implique $f$. Un *impliquant premier* de $f$ est un impliquant $t$ de $f$ tel qu'aucun sous-ensemble de $t$ n'est un impliquant de $f$. Une instance partielle est un tuple $\boldsymbol{y} \in \{0,1,\bot\}^n$. Intuitivement, si $y[i] = \bot$, alors la valeur de la i-ème feature est indéfinie. $\mathbf{Comp(y)}$ désigne l'ensemble des complétions de $\boldsymbol{y}$ . Nous disons que $\boldsymbol{y}$ est subsumé par $\boldsymbol{x}$ s'il est possible

d'obtenir $y$ à partir de $\boldsymbol{x}$ en échangeant certaines valeurs indéfinies par des valeurs de $\boldsymbol{x}$, noté $\boldsymbol{x} \subseteq \boldsymbol{y}$, nous définissons $|\boldsymbol{y}|_{\perp} = |\{i \in \{1, \ldots, n\}, |\boldsymbol{y}[i] = \perp\}|$.

## 2.2 Arbre de décision

**Arbre de décision binaire.** Un arbre de décision binaire sur $X_n$ est un arbre binaire $T$, dont chacun des nœuds internes est étiqueté avec l'une des $n$ variables booléennes d'entrée de $X_n$, et dont les feuilles sont étiquetées par 0 ou 1. Il est supposé que chaque variable apparaît au plus une fois sur n'importe quel chemin racine-feuille (propriété de read-one). La valeur $T(\boldsymbol{x}) \in \{0, 1\}$ de $T$ sur une instance d'entrée $\boldsymbol{x}$ est donnée par l'étiquette de la feuille atteinte depuis la racine comme suit à chaque nœud. La taille de $T$, notée $|T|$, est donnée par le nombre de ses nœuds. La classe des arbres de décision est notée $\texttt{DT}_n$.

Il est bien connu que tout arbre de décision $T \in \texttt{DT}_n$ peut être transformé en temps linéaire en une disjonction de termes équivalente, notée $\texttt{DNF}(T)$. Cette DNF est une DNF orthogonale, dans laquelle chaque terme correspond à un chemin de la racine à une feuille étiquetée 1. $T$ peut être transformé en une conjonction de clauses, notée $\texttt{CNF}(T)$ (Audemard et al., 2022c).

## 2.3 DNF orthogonale

Un problème classique de la théorie booléenne est de dériver une forme normale disjonctive orthogonale d'une fonction booléenne arbitraire. Pour cela, considérons la DNF :

$$\phi = \bigvee_{k=1}^{m} \left( \bigwedge_{i \in A_k} x_i \bigwedge_{j \in B_k} \overline{x_j} \right) \tag{1}$$

Où $A_k \cap B_k = \emptyset$ et pour tout $k = 1, 2, \ldots, m$. $C_k = \left( \bigwedge_{i \in A_k} x_i \bigwedge_{j \in B_k} \overline{x_j} \right)$ est le $k$-ème terme de la DNF.

**Définition 1** (DNF orthogonale)**.** *Une* DNF *de la forme 1 est dite orthogonale, si* $(A_k \cap B_l) \cup (A_l \cap B_k) \neq \emptyset$ *pour tout* $k, l \in \{1, 2, \ldots, m\}$ *et* $k \neq l$.

**Proposition 1.** *Soit une* DNF $\phi$ *de la forme 1, alors le nombre de ses modèles est égal à :*

$$w(\phi) = \sum_{k=1}^{m} 2^{n - |A_k| - |B_k|} = \sum_{k=1}^{m} \alpha_k$$

*où* $\alpha_k = 2^{n - |A_k| - |B_k|}$ *pour chaque terme* $C_k$ *de la formule* DNF $\phi$.

# 3 Explications probabilistes

## 3.1 Explications abductives

Les arbres de décision sont explicables localement, par construction, toute instance d'entrée $\boldsymbol{x}$ est associée à un unique chemin de la racine à la feuille dans un arbre de décision qui

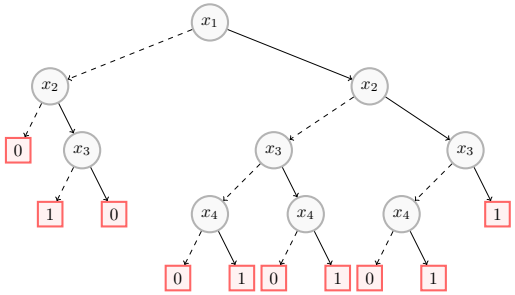


FIG. 1 – Un arbre de décision $T$ sur les attributs $\{x_1, x_2, x_3, x_4\}$.

conduit à une décision, Une raison directe ou explication de chemin (Izza et al., 2020) pour $\boldsymbol{x}$ étant donné $T$ est un terme de la $\texttt{DNF}(T)$, noté $p_{\boldsymbol{x}}^T$, correspondant au chemin unique de la racine à la feuille de $T$. Cette raison directe est spécifique aux arbres de décision et également aux forêts aléatoires (Audemard et al., 2022d,a,b). Une autre notion importante pour expliquer les prédictions et qui n'est pas spécifique aux arbres de décision est la suivante :

**Définition 2** (Raison suffisante)**.** *Soit $f \in \mathcal{F}_n$ et $\boldsymbol{x} \in \{0,1\}^n$ tel que $f(\boldsymbol{x}) = 1$. Une « raison suffisante » pour $\boldsymbol{x}$ étant donné $f$ est un impliquant premier $t$ de $f$ qui couvre $\boldsymbol{x}$.*

Nous rappelons que pour la classe $\texttt{DT}_n$, des raisons suffisantes peuvent être trouvées en temps polynomial, voir (Audemard et al., 2022c), ainsi qu'une *raison suffisante de taille minimale* pour $x$ donné $f$ est une raison suffisante qui contient un nombre minimal de littéraux.

**Exemple 1.** *La $\texttt{DNF}$ représentant l'arbre de la figure 1 est $\phi = (x_1 \wedge x_2 \wedge x_3) \vee (\overline{x_1} \wedge x_2 \wedge \overline{x_3}) \vee (x_1 \wedge x_2 \wedge \overline{x_3} \wedge x_4) \vee (x_1 \wedge \overline{x_2} \wedge x_3 \wedge x_4) \vee (x_1 \wedge \overline{x_2} \wedge \overline{x_3} \wedge x_4)$. et $w(T) = w(\phi) = 2^1 + 2^1 + 2^0 + 2^0 + 2^0 = 4 + 3 = 7$. Pour l'instance $\boldsymbol{x} = (1,1,1,1)$, nous observons que $T(\boldsymbol{x}) = 1$. La raison directe de $\boldsymbol{x}$ étant donné $T$ est $p_{\boldsymbol{x}}^T = x_1 \wedge x_2 \wedge x_3$, $x_1 \wedge x_4$ et $x_1 \wedge x_2 \wedge x_3$ sont deux raisons suffisantes de $\boldsymbol{x}$. $x_1 \wedge x_4$ est l'unique raison suffisante de taille minimale.*

## 3.2 Explications $\delta$-probable

La notion de raison suffisante est souvent considérée comme un concept naturel d'explication du résultat d'un classifieur, mais elle impose une restriction stricte en exigeant que toutes les complétions d'une instance partielle soient classées de la même manière et ces raisons peuvent être aussi d'une grande taille. Pour assouplir cette limitation, une généralisation probabiliste des explications a été proposée par (Wäldchen et al., 2021).

**Définition 3** (raison $\delta$-probable)**.** *Soit $\delta \in (0,1]$, une raison $\delta$-probable pour $\boldsymbol{x}$ étant donné une fonction booléenne $f \in F_n$ tel que $f(\boldsymbol{x}) = 1$ est un terme partiel $t_y$ tel que $t_y \subseteq t_x$ et :*

$$\mathbb{P}_z[f(z) \mid t \subseteq t_z] = \frac{|\{z \in \mathbf{Comp}(\mathbf{y}) \mid f(z) = f(\boldsymbol{x})\}|}{2^{|y|_\perp}} = \frac{h(\boldsymbol{y})}{2^{|y|_\perp}} \geq \delta \tag{2}$$

Dans le cas où $f$ est représentée par un arbre de décision $T$, l'équation 2 peut être réécrite :

$$\frac{w(\phi \wedge t_y)}{2^{|y|_\perp}} \geq \delta \tag{3}$$

$\phi = \texttt{DNF}(T)$ et $t_y = \bigwedge_{i=1}^{n} x_i^{y_i}$, $|y|_{\perp} = n - |y|$ et $w(\phi \wedge t_y)$ est le nombre de modéle de $\phi \wedge t_y$.

**Exemple 2.** *Soit $T$ l'arbre de la figure 1, et l'instance $\boldsymbol{x} = (1, 1, 1, 1)$ ($T(\boldsymbol{x}) = 1$). On observe que $t_{\{x_1,x_2\}} = x_1 \wedge x_2$ et $t_{\{x_1,x_3\}} = x_1 \wedge x_3$ sont des raisons $\frac{3}{4}$-probable pour $\boldsymbol{x}$ étant donné $T$ et $t_{\{x_1,x_4\}} = x_1 \wedge x_4$ est une raison 1-probable (également une raison suffisante).*

# 4 Algorithmes gloutons

Le problème de décision vérifiant si $\boldsymbol{x}$ admet une raison $\delta$-probable de certaine taille $k$ pour une fonction booléenne $f$, où $f$ est présentée par une formule CNF, est $\mathsf{NP}^{\mathsf{PP}}$-complet (Wäldchen et al., 2019), et $\mathsf{NP}$-difficile lorsque $f$ est un arbre de décision (Arenas et al., 2022). Un encodage SAT a été proposé par (Arenas et al., 2022) pour dériver une raison $\delta$-probable pour $\boldsymbol{x}$ étant donné $T$. Cependant, le temps requis pour obtenir des résultats en utilisant cet méthode est élevé dans de nombreuses cas comme souligné dans (Bounia et Koriche, 2023).

## 4.1 Algorithme glouton

Dans ce qui suit, nous proposons un algorithme glouton pour dériver un ensemble $S$ basé sur un ensemble d'attributs $E \subseteq \boldsymbol{x}$ d'une taille exactement $k \leq |E|$ (ou d'une taille au plus $k$) et un paramètre de confiance $\delta \in (0, 1]$ à déterminer. Pour un arbre de décision $T \in \texttt{DT}_n$, on note $\phi = \texttt{DNF}(T)$, on exploit le fait que cette DNF est orthogonale (Audemard et al., 2022c) pour effectuer les calculs de la manière la moins coûteuse possible. L'orthogonalité de $\phi$ permet de compter les modèles de $T$ en temps linéaire. Ainsi, la vérification de l'inégalité $\frac{w(\phi \wedge t_s)}{2^{n-|s|}} \geq \delta$ peut être effectuée en temps linéaire. Dans la suite, nous détaillons explicitement notre algorithme glouton.

**Algorithm 1**

**Input:** un arbre $T$, un ensemble $E \subseteq \boldsymbol{x}$, et $k \leq |E|$
**Output:** une raison $\delta^*$-probable
$E \leftarrow Lit(\boldsymbol{x})$, $S \leftarrow \emptyset$, $\phi \leftarrow \texttt{DNF}(T)$ ; `/*E peut être l'instance` $\boldsymbol{x}$ `ou bien` $p^T_{\boldsymbol{x}}$, `ou bien un sous-ensemble` $I \subseteq \boldsymbol{x}$ `*/`

**for** $l \in \{1, ..., k\}$ **do**
  $e^* \longleftarrow \underset{c \in E}{\operatorname{argmax}}\, h(S \cup \{e\})$
  $S \longleftarrow S \cup \{e^*\}$
  $E \longleftarrow E - \{e^*\}$
$\delta^* = \frac{w(\phi \wedge t_S)}{2^{n-k}}$
**return** $S, \delta^*$

L'algorithme 1 est une version adaptée de l'algorithme GA proposé dans (Bounia et Koriche, 2023). Cet algorithme vise à trouver une explication probabiliste de taille $k$ (ou au plus $k$) qui minimise l'erreur de classification (maximise la valeur de $\delta$) (Bounia et Koriche, 2023).

**Proposition 2.** *Soit un arbre de décision $T \in \texttt{DT}$ et une instance $\boldsymbol{x}$, l'algorithme 1 s'exécute en temps $O(k \cdot n^2|T|)$.*

**Exemple 3.** *Pour l'arbre de la figure 1. Soit* $\boldsymbol{x} = (1, 1, 1, 1)$. *Nous cherchons une raison probable d'une taille* $k = 2$. *Les étapes de l'algorithme 1 :* $x_1 = \underset{e \in \{x_1, x_2, x_3, x_4\}}{\operatorname{argmax}} h(\{e\})$, *alors* $S = \{x_1\}$ *et* $x_4 = \underset{e \in \{x_2, x_3, x_4\}}{\operatorname{argmax}} h(\{x_1\} \cup \{e\})$, *On obtient* $S = \{x_1, x_4\}$ *et* $\frac{w(\phi \wedge x_1 \wedge x_4)}{2^{4-2}} = 1$. *Alors l'algorithme 1 a capturé l'unique raison suffisante de taille minimale.*

### 4.2 Dériver une raison $\delta$-probable

Finalement, étant donné que l'algorithme 1 s'exécute en temps **linéaire** et capture en pratique le plus souvent une explication probabiliste avec un paramètre $\delta$ fiable (de valeur maximale). Dans la suite, nous allons légèrement modifier l'algorithme 1 afin de dériver une explication probabiliste pour l'instance $\boldsymbol{x}$ étant donné $T$ et un paramètre de confiance $\delta \in (0, 1]$. Est-il également possible de calculer une explication probabiliste spécifique basée sur un sous-ensemble de littéraux $E \subseteq \boldsymbol{x}$. Pour cela, nous ajustons l'entrée de l'algorithme 2 en utilisant le sous-ensemble $E$ comme l'entrée de l'algorithme 2.

**Algorithm 2** Dérivation d'une raison $\delta$-probable

**Input:** un arbre de décision $T$, $\delta \in (0, 1]$, un sous-ensemble $E \subseteq \boldsymbol{x}$
**Output:** une raison $\delta-$probable

$\phi \leftarrow \texttt{DNF}(T)$, $S \leftarrow \emptyset$

**for** $l \in \{1, ..., |E|\}$ **do**
  $e^* \longleftarrow \underset{c \in E}{\operatorname{argmax}}\, h(S \cup \{e\})$
  $S \longleftarrow S \cup \{e^*\}$
  **if** $\frac{h(S)}{2^{n-l}} \geq \delta$ **then**
    **break**
    **return** $S$
  $E \longleftarrow E - \{e^*\}$

$\delta^* = \frac{w(\phi \wedge t_S)}{2^{n-|S|}}$

**return** $S, \delta^*$

# 5 Expérimentations

Nous avons mené plusieurs expériences pour évaluer la performance de nos deux algorithmes. Nos objectifs sont les suivants :

— Mesurer l'exactitude des résultats de l'algorithme 1 pour calculer une raison probable d'une certaine taille $k$ et d'une taille au plus $k$. En particulier, nous avons comparé la valeur de $\delta^*$ associée à la raison trouvée par l'algorithme 1 à la valeur optimale $\delta_{opt}$ obtenue grâce à la recherche dichotomique de l'encodage SAT (Arenas et al., 2022).

| dataset | | | decision tree | | | \|Reason\| | | | $\lvert\delta_{opt} - \delta^*_{algo_1}\rvert$ | | | | SAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| name | #F | #I | %A | $\lvert T\rvert$ | Depth | $\lvert p_x^T\rvert$ | $\lvert SR\rvert$ | $\lvert MR\rvert$ | k | $t_x$ | $p_x^T$ | $SR$ | Time (s) |
| horse | 29 | 299 | 84.44 | 34 | 13 | 6.0 | 6.6 | 5.2 | 5 | 0.0085 | 0.0004 | 0.0004 | 20,45 |
| hungarian | 13 | 294 | 68.54 | 62 | 12 | 6.0 | 5.8 | 4.9 | 5 | 0.008 | 0.002 | 0.002 | 5.53 |
| primary. t | 23 | 399 | 87.25 | 53 | 14 | 6.0 | 7.1 | 4.8 | 5 | 0.0294 | 0.0004 | 0.0004 | 17.25 |
| mushroom | 17 | 8124 | 100.0 | 20 | 7 | 5.0 | 4.7 | 4.4 | 5 | 0.0002 | 0.0002 | 0.0002 | 2.53 |
| cars | 21 | 406 | 97.54 | 30 | 10 | 4.4 | 5.5 | 4 | 4 | 0.0506 | 0.0026 | 0.0026 | 9.46 |
| glass | 31 | 214 | 83.08 | 36 | 11 | 6.9 | 8.4 | 6.4 | 5 | 0.009 | 0.0047 | 0.0051 | 3.82 |
| placement | 18 | 215 | 95.38 | 19 | 10 | 4.0 | 4.4 | 3.2 | 3 | 0.0001 | 0.0001 | 0.0001 | 0.72 |
| spect | 19 | 265 | 78.75 | 42 | 15 | 6.3 | 6.3 | 4.7 | 5 | 0.04 | 0.002 | 0.002 | 3.71 |
| colic | 40 | 368 | 80.18 | 55 | 13 | 8.0 | 10.2 | 7.5 | 6 | 0.0003 | 0.0002 | 0.0004 | 19.23 |
| biomed | 15 | 209 | 98.41 | 21 | 11 | 4.6 | 4.1 | 3.7 | 4 | 0.0004 | 0.0001 | 0.0001 | 0.67 |
| student-por | 30 | 649 | 91.79 | 33 | 9 | 5.0 | 6.3 | 4.9 | 4 | 0.0037 | 0.0007 | 0.0007 | 41.67 |
| tic-tac-toe | 9 | 958 | 97.92 | 83 | 9 | 5.8 | 4.8 | 4.4 | 4 | 0.0062 | 0.0005 | 0.0001 | 1.22 |
| schizo | 33 | 340 | 93.14 | 34 | 11 | 5.9 | 5.3 | 4.7 | 5 | 0.0019 | 0.0002 | 0.0002 | 4.53 |
| vehicle | 23 | 846 | 96.06 | 31 | 12 | 5.7 | 6.6 | 5.2 | 5 | 0.0005 | 0.0007 | 0.0006 | 3.34 |
| balance | 17 | 625 | 86.7 | 77 | 12 | 5.6 | 5.8 | 4.7 | 5 | 0.0048 | 0.0005 | 0.0005 | 13.17 |
| compas | 40 | 6172 | 66.14 | 570 | 20 | 11.1 | 9.4 | 6.6 | 7 | 0.006 | 0.006 | 0.006 | 1482.32 |
| employee | 63 | 4653 | 82.45 | 653 | 20 | 10.4 | 12.0 | 8.1 | 7 | 0.13 | 0.08 | 0.06 | 1314.84 |
| fetal. h | 93 | 2127 | 93.42 | 110 | 19 | 12.2 | 17.3 | 10.8 | 7 | 0.18 | 0.12 | 0.14 | 1125.83 |

TAB. 1 – Statistiques sur la fiabilité des explications probabilistes générées par l'algorithme 1 et comparaison avec la méthode SAT

— Évaluer le gain de l'intelligibilité résultant de l'accent mis sur la taille des explications probabilistes (calculée à l'aide de l'algorithme 2) par rapport à la taille des raisons directes, suffisantes (et de taille minimale) d'une instance donnée. Nous avons constaté que l'algorithme 1 pour calculer des raisons $\delta$-probable est plus efficace en terme de temps de calcul que la méthode SAT et peut traiter des problèmes à plus grande échelle là où la méthode SAT devient inefficace en termes de temps de calcul.

## 5.1 Protocole Expérimental

Nous avons considéré 32 datasets, qui sont des références standard provenant des célèbres sites Kaggle [1], OpenML [2] et UCI [3]. Notamment, mnist38 et mnist49 sont des sous-ensembles de l'ensemble de données mnist. Les attributs catégoriels ont été traitées comme des nombres arbitraires. Quant aux attributs numériques, ces attributs ont été binarisés à l'aide de l'algorithme d'apprentissage d'arbre de décision utilisé. Les performances de classification pour $T_b$ ont été mesurées comme la précision moyenne obtenue sur un ensemble de test de plus de $150$ instances. Pour l'apprentissage d'arbre de décision, nous avons utilisé l'algorithme CART, et plus précisément son implémentation fournie par la bibliothèque Scikit-Learn. (Pedregosa et al., 2011). Tous les hyperparamètres ont été réglés sur leurs valeur par défaut.

Pour chaque ensemble de données $b$, chaque arbre de décision $T_b$, et chaque instance $\boldsymbol{x}$ de l'ensemble de test correspondant, dans le but de tester la fiabilité de notre algorithme, nous avons calculé le $\delta^*$ qui correspond à une raison probable de taille au plus $k$. Pour ce faire, nous avons commencé par utiliser l'instance $E = t_{\boldsymbol{x}}$, puis la raison directe $E = p^T_{\boldsymbol{x}}$, et enfin une raison suffisante $E = SR(\boldsymbol{x})$, que nous avons comparé au $\delta_{opt}$ (voir tableau 1).

Pour calculer une $\delta_{opt}$ qui correspond à une raison probable de taille exacte $k$, nous avons effectué une recherche dichotomique (binaire) en étendant l'encodage `CNF` proposé par (Arenas et al., 2022) en ajoutant l'encodage `CNF` de la contrainte de cardinalité $(\sum_{i\in E} x_i) = k$.

1. `www.kaggle.com`
2. `www.openml.org`
3. `archive.ics.uci.edu/ml/`

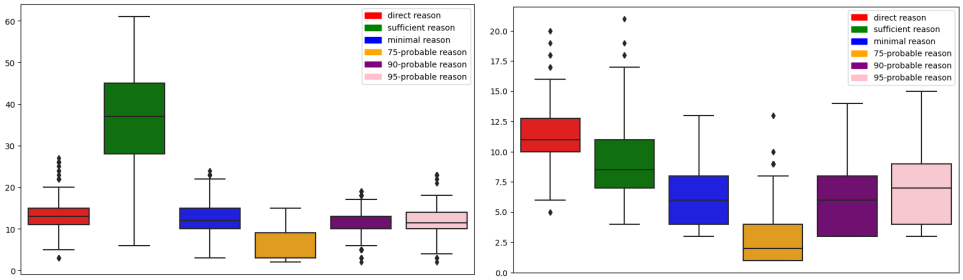


FIG. 2 – Boîtes à moustaches pour "spambase" (à gauche) et "compas" (à droite), représentant les tailles des raisons directes, des raisons suffisantes, des raisons suffisantes de taille minimale, et des raisons $75\%$, $90\%$, $95\%$-probable.

Concernant le $\delta_{opt}$ qui correspond a une raison probable de taille au plus $k$, l'encodage SAT est également étendu (Arenas et al., 2022) en ajoutant la clause $C_E = \{x_i : (x_E)_i = 1\} \vee \{\overline{x_i} : (x_E)_i = 0\}$ à la CNF de l'encodage original. La recherche dichotomique a été effectué avec une précision d'erreur de $10^{-3}$ et un temps limite de $1800(s)$, ce qui équivaut à faire 10 appels SAT. En ce qui concerne le tableau 2, il est à noter que l'encodage SAT ne permet pas de passer à l'échelle en raison de l'explosion de la consommation de mémoire. Par conséquent, pour la comparaison (pour $k = 7$), nous nous sommes limités à partir de la raison directe. Pour le calcul de $\delta_{opt}$, nous avons utilisé un algorithme exact consistant à tester toutes les combinaisons $\frac{h(S)}{2^{n-7}}$ de sous-ensembles $S$ de $p_x^T$ de taille exacte 7, et prendre la valeur maximale.

Dans le but d'évaluer l'amélioration de l'intelligibilité résultant de l'accent mis sur la taille des explications probabilistes, nous avons rapporté les tailles d'une raison directe, une raison suffisante pour $\boldsymbol{x}$ donnée $T_b$, une raison suffisante de taille minimale pour $\boldsymbol{x}$ étant donné $T_b$ est calculée à l'aide de l'outil PyXAI (Audemard et al., 2023). Nous avons ensuite rapporté les temps de calcul requis pour reporter $\delta_{opt}$ en utilisant la recherche dichotomique de l'encodage SAT. Pour cela, nous avons utilisé la bibliothèque $Pysat$ [4], qui fournit l'implémentation du solveur GLUCOSE 4 (nous avons défini une limite de temps de 1800 secondes). Et pour calculer des raisons suffisantes de taille minimale, nous avons utilisé la bibliothèque Pysat, qui permet d'utiliser le solveur PARTIAL MAXSAT $RC2$. Ce solveur a été exécuté en utilisant les paramètres correspondant à la configuration "Glucose".

Toutes les expériences ont été menées sur un ordinateur équipé d'un processeur Intel(R) Core(TM) $i9 - 9900$ @ $3,10$ GHz et de $64$ GiB de mémoire.

## 5.2 Résultats

Le tableau 1 présente un extrait de nos résultats pour 18 datasets. La première colonne donne le nom du dataset $b$. $F$ représente le nombre d'attributs binaires, $I$ le nombre d'instances, $\%A$ la précision de l'arbre $T_b$, $|T|$ sa taille, et $Depth$ représente la profondeur de l'arbre. La colonne $|Reason|$ indique la taille moyenne de différentes raisons calculées : $P_x^T$, $|SR|$, $|MR|$ représentent respectivement la taille de la raison directe, la taille de la raison suffisante, et la taille de la raison suffisante minimale. $|\delta_{opt} - \delta_{algo_1}|$ indique l'erreur moyenne de $\delta$

4. https://pysathq.github.io/

correspondant à la raison que l'algorithme 1 retourne pour une taille au plus $k$, en partant respectivement de l'instance complète $|t_x|$, de la raison directe $|p_x^T|$, et de la raison suffisante $|SR|$. **SAT** indique les temps de calcul moyens de la recherche dichotomique en partant de la raison directe. Les ensembles de données en magenta indiquent que le temps limite de $1800$ secondes a été atteint au moins une fois. Nous remarquons que l'erreur moyenne $|\delta_{opt} - \delta_{algo_1}|$ est en général de l'ordre de $10^{-3}$, en particulier lorsque l'entrée de l'algorithme est $p_x^T$ et $SR$. Cependant, la précision diminue légèrement lorsque l'entrée est l'instance complète $x$. Cela montre que notre algorithme glouton trouve généralement une raison probable de taille au plus $k$ avec le paramètre de confiance $\delta$ le plus élevé possible. Nous notons également que l'erreur $|\delta_{opt} - \delta_{algo_1}|$ est élevée pour les ensembles de données où le temps limite est dépassé (en magenta). Cela est dû au fait que la recherche dichotomique s'est arrêtée avant d'atteindre une précision de $\epsilon = 10^{-3}$. Dans le but d'améliorer l'intelligibilité de nos explications, nous avons calculé, à l'aide de l'algorithme 2 (figure 2), des raisons $\delta$-probables pour $\delta = 0,75$, $\delta = 0,9$ et $\delta = 0,95$. Nos résultats montrent que les explications probabilistes sont généralement plus concises que les explications abductives, y compris celles de taille minimale ("raisons suffisantes de taille minimale").

En ce qui concerne le temps de calcul requis par la recherche dichotomique de l'encodage SAT, nous avons remarqué que ce temps était très élevé, atteignant jusqu'à $25$ minutes dans certains cas, notamment lorsque la taille de l'arbre $|T|$ est grande (comme dans le cas de "Compas" et "Employee") ou lorsque le nombre d'attributs binaires est élevé (comme dans le cas de "Fetal. h"). Nous n'avons pas inclus les temps de calcul de notre algorithme glouton, car le temps moyen par instance nécessaire pour toutes nos expériences ne dépasse pas $0,5$ seconde, ce qui démontre l'efficacité computationnelle de notre algorithme par rapport à l'encodage SAT. Il est intéressant de noter que les raisons probables d'une certaine taille $k$ (ou au plus $k$) offrent aux utilisateurs la possibilité de contrôler la taille de l'explication en se basant sur un sous-ensemble de variables de leur choix. Il est également essentiel de souligner que ces raisons sont bien plus pertinentes pour l'utilisateur qu'une raison sélectionnée de manière aléatoire.

Les résultats de nos expérimentations du tableau 2 mettent en évidence la difficulté du calcul des raisons probabilistes avec l'encodage SAT, le temps limite de $1800$ secondes étant systématiquement atteint. Nous avons inclus des ensembles de données de grandes dimensions, pour lesquels l'encodage SAT ne parvient pas à passer à l'échelle. Nous avons utilisé la raison directe $p_x^T$ comme entrée de l'algorithme 1 et avons fixé la taille de la sortie à $k = 7$. Nous remarquons que l'algorithme 1 est très efficace sur ces ensembles de données de grandes dimensions, et l'erreur ne dépasse pas l'ordre de $10^{-3}$. Cela confirme la fiabilité de nos résultats, comme présentés dans le tableau 1. Il est également à noter que cela confère un avantage à notre approche, en particulier lorsque l'encodage SAT ne parvient pas à passer à l'échelle.

Afin d'illustrer le gain d'intelligibilité obtenus lors de la transition des explications abductives (raisons directes et raisons suffisantes) vers des raisons probabilistes (raisons $\delta$-probables) pour $150$ instances, nous avons créé plusieurs diagrammes en boîte pour deux datasets : "compas" (à droite), qui illustre la transition des raisons directes vers des raisons $\{75\%, 90\%, 95\%\}$-probables, et "spambase" (à gauche), qui illustre la transition des raisons suffisantes vers des raisons $\{75\%, 90\%, 95\%\}$-probables. La figure 2 présente ces diagrammes en boîte. Nous pouvons observer qu'une réduction significative du nombre d'attributs utilisés dans les raisons

| Dataset | #I | #F | %A | $\lvert\delta_{opt} - \delta^*\rvert$ | $\lvert P_x^T \rvert$ |
|---|---|---|---|---|---|
| Gisette | 5000 | 7000 | 98.56 | 0 | 21.42 |
| Mnist38 | 13966 | 784 | 95.44 | 0.0003 | 17.89 |
| Mnist49 | 13782 | 784 | 95.48 | 0 | 15.57 |
| Christine | 1636 | 5418 | 61.25 | 0.001 | 9.47 |
| Bank | 41188 | 882 | 89.49 | 0.0003 | 13 |
| Dexter | 20000 | 600 | 90.70 | 0 | 8.32 |
| Gina-agnostic | 970 | 48842 | 85.84 | 0.0001 | 9.84 |
| Gina | 970 | 3468 | 84.53 | 0.0001 | 9.69 |
| Farm-ads | 54877 | 1543 | 86.8 | 0.0003 | 23.15 |
| Cnae | 1080 | 856 | 92.59 | 0.0006 | 19.07 |
| Dorothea | 1150 | $10^5$ | 91.8 | 0 | 12.9 |
| Adult | 48842 | 2974 | 81.16 | 0.001 | 16.43 |
| Spambase | 4601 | 236 | 92.11 | 0.0002 | 16.09 |
| Ad-data | 5000 | 1023 | 99.19 | 0 | 9.29 |

TAB. 2 – Tableau des résultats pour 14 datasets : nombre d'instances $I$, nombre d'attributs binaires $F$, précision $\%A$ et erreur moyenne de $|\delta_{opt} - \delta^*|$ pour $k = 7$.

directes peut se produire lors de la transition vers une raison $0,75$-probable, de même que pour les raisons suffisantes.

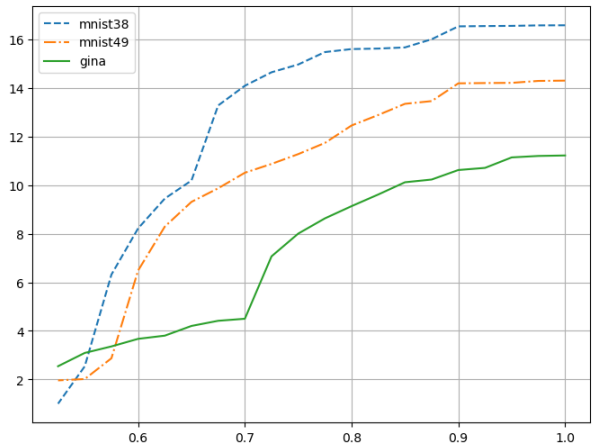


FIG. 3 – Taille moyenne des raisons $\delta$-probables lorsque $\delta$ varie de $0,5$ à $1$ (raisons suffisantes) pour les ensembles de données "mnist38", "mnist49" et "gina".

Enfin, étant donné que la réduction de la taille des raisons obtenue en considérant des raisons $\delta$-probables de $75\%$ par rapport aux raisons directes et aux raisons suffisantes semblait significative, nous avons également réalisé des expériences supplémentaires pour obtenir une vision plus claire de la réduction qui peut être obtenue avec des variations de $\delta$. Nous avons calculé les tailles des raisons $\delta$-probables pour les instances alors que $\delta$ varie de $0.5$ à $1$. La figure présente de tels graphiques pour les ensembles de données "mnist38", "mnist49"

et "gina". Comme prévu, on peut observer que la taille moyenne des raisons $\delta$-probables augmente progressivement lorsque $\delta$ augmente et se stabilise lorsque l'algorithme capture une raison suffisante. Ce qui montre clairement le gain de l'intelligibilité obtenue.

# 6 Conclusion

Dans cet article, nous avons exploité les résultats des travaux récents sur l'approximation des explications probabilistes afin d'améliorer l'intelligibilité, en particulier dans le contexte des arbres de décision. Par nature, les explications probabilistes ne peuvent pas être plus grandes que les raisons suffisantes, mais elles se révèlent être des concepts précieux pour obtenir des explications plus compréhensibles pour les utilisateurs humains. Toutes ces raisons sont plus petites que les instances elles-mêmes. Bien que les raisons suffisantes de taille minimale soient les explications abductives les plus courtes possibles. Nos expériences ont montré qu'une réduction supplémentaire de taille peut être obtenue avec les raisons $\delta$-probables. De plus, nous constatons que notre algorithme glouton permet de dériver des raisons probables fiables et concises, avec un coût computationnel remarquablement bas par rapport à la méthode SAT. Nous pouvons affirmer que la dérivation de raisons probables fiables est considérablement simplifiée en utilisant notre algorithme glouton, rendant le gain en intelligibilité qu'elles apportent presque gratuit. Dans nos futurs travaux, nous prévoyons d'étudier l'approximation des explications probabilistes sur d'autres types de classifieurs, à savoir les forêts aléatoires et les arbres boostés.

# Références

Arenas, M., P. Barceló, M. Romero Orth, et B. Subercaseaux (2022). On computing probabilistic explanations for decision trees. *Advances in Neural Information Processing Systems 35*, 28695–28707.

Audemard, G., S. Bellart, L. Bounia, F. Koriche, J. Lagniez, et P. Marquis (2022a). On preferred abductive explanations for decision trees and random forests. In *Proc. of IJCAI'22*.

Audemard, G., S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, et P. Marquis (2022b). Les raisons majoritaires : des explications abductives pour les forêts aléatoires. *EGC'2022 38*.

Audemard, G., S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, et P. Marquis (2022c). On the explanatory power of boolean decision trees. *Data  Knowledge Engineering 142*, 102088.

Audemard, G., S. Bellart, L. Bounia, F. Koriche, J.-M. Lagniez, et P. Marquis (2022d). Trading complexity for sparsity in random forest explanations. In *Proc. of AAAI'22*.

Audemard, G., S. Bellart, L. Bounia, J.-M. Lagniez, P. Marquis, et N. Szczepanski (2023). Pyxai : calculer des explications pour des modèles d'apprentissage supervisé. *EGC*.

Bounia, L. et F. Koriche (2023). Approximating probabilistic explanations via supermodular minimization (corrected version). In *Uncertainty in Artificial Intelligence (UAI 2023).*, Volume 216, pp. 216–225.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Ignatiev, A., N. Narodytska, et J. Marques-Silva (2019). Abduction-based explanations for machine learning models. In *Proc. of AAAI'19*, pp. 1511–1519.

Izza, Y., A. Ignatiev, et J. Marques-Silva (2020). On explaining decision trees. *ArXiv abs/2010.11034*.

Lundberg, S. et S.-I. Lee (2017). A unified approach to interpreting m(ijcaiodel predictions. In *Proc. of NIPS'17*, pp. 4765–4774.

Miller, G. A. (1956). The magical number seven, plus or minus two : Some limits on our capacity for processing information. *The Psychological Review 63*(2), 81–97.

Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence 267*, 1–38.

Molnar, C. (2020). *Interpretable Machine Learning*. Leanpub.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, et E. Duchesnay (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Ribeiro, M. T., S. Singh, et C. Guestrin (2016). "why should I trust you ?" : Explaining the predictions of any classifier. In *Proc. of SIGKDD'16*, pp. 1135–1144.

Ribeiro, M. T., S. Singh, et C. Guestrin (2018). Anchors : High-precision model-agnostic explanations. In *Proc. of AAAI'18*, pp. 1527–1535.

Shih, A., A. Choi, et A. Darwiche (2018). A symbolic approach to explaining bayesian network classifiers. In *Proc. of IJCAI'18*, pp. 5103–5111.

Wäldchen, S., J. MacDonald, S. Hauch, et G. Kutyniok (2019). The computational complexity of understanding network decisions. *CoRR abs/1905.09163*.

Wäldchen, S., J. Macdonald, S. Hauch, et G. Kutyniok (2021). The computational complexity of understanding binary classifier decisions. *J. Artif. Intell. Res. 70*, 351–387.

# Summary

This work deals with explainable artificial intelligence (XAI), specifically focusing on improving the intelligibility of decision trees through reliable and concise probabilistic explanations. Decision trees are popular because they are considered highly interpretable. Due to cognitive limitations, abductive explanations can be too large to be interpretable by human users. When this happens, decision trees are far from being easily interpretable. In this context, our goal is to enhance the intelligibility of decision trees by using probabilistic explanations. Drawing inspiration from previous work on approximating probabilistic explanations, we propose a greedy algorithm that enables us to derive concise and reliable probabilistic explanations for decision trees. We provide a detailed description of this algorithm and compare it to the state-of-the-art SAT encoding, emphasizing the gains in intelligibility and highlighting its empirical effectiveness.